Supplemental Figures

Fig. S1

Fig. S1 | Mutational profiles (rainfall plots) for all 97 breast tumors, related to Fig. 1. Each panel represents a tumor with the X-axis showing mutations ordered by mutation number (from the first mutated position on chromosome 1 to last mutated position on chromosome X) and the Y-axis represents intermutation distance, in log-scale. This figure is provided as an additional PDF file.

Fig. S2

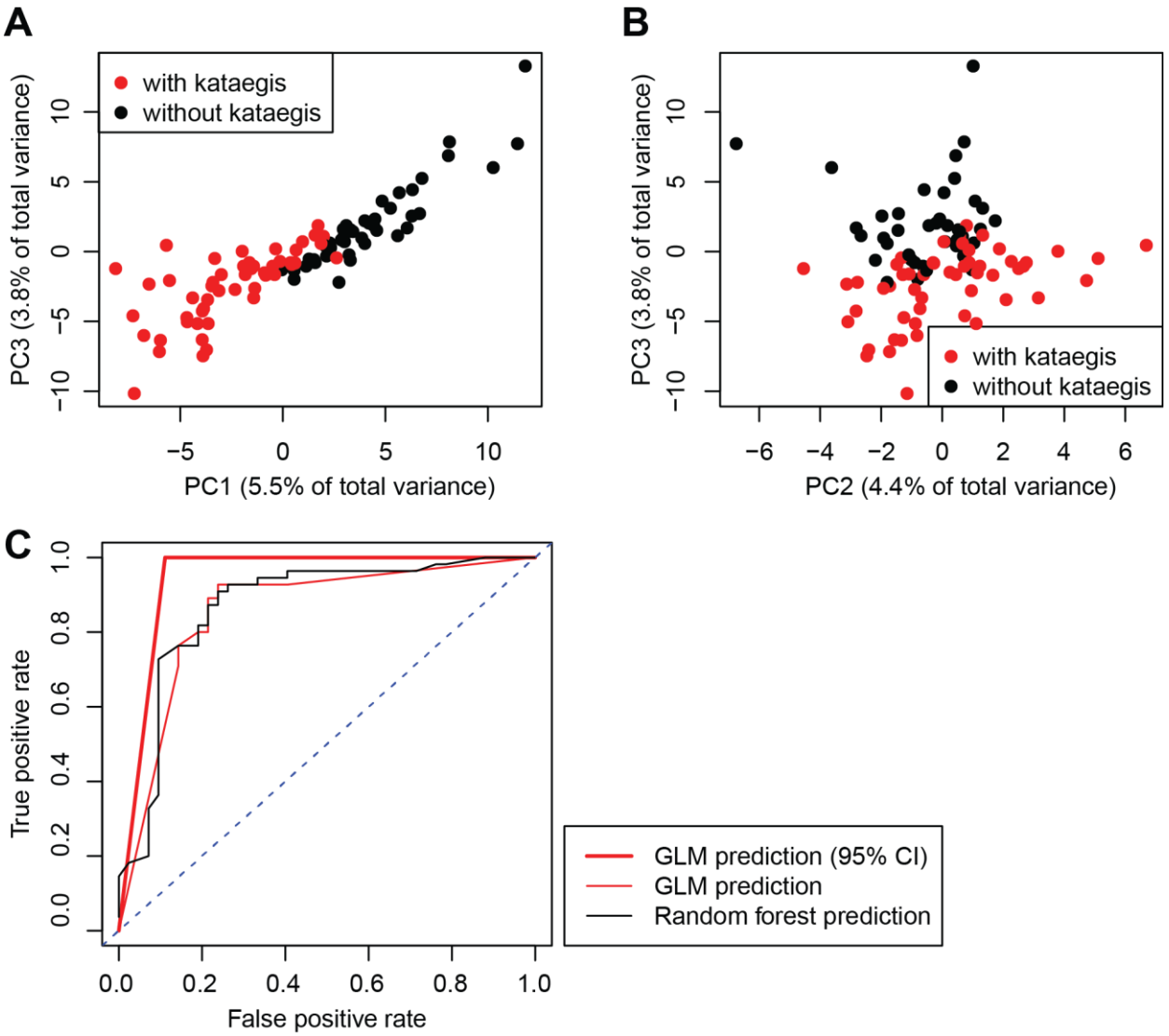


Fig. S2 | Gene expression differences between tumors with and without kataegis and prediction of kataegis in tumors of unknown status, related to Fig. 4. PCA was performed on the 628 genes with significant expression differences between 55 tumors with and 42 tumors without kataegis. Scatterplots showing **(A)** principal components 1 and 3, and **(B)** principal components 2 and 3. The associations between PC1 and PC2 is shown in Fig. 4A. (**C**) ROC curves were drawn for three prediction methods: 1) GLM prediction; 2) GLM prediction using 95% CI values from the GLM standard errors; and 3) a random forest model. The figure shows that GLM prediction and random forest have similar ROC AUC (0.84 and 0.87, respectively), while using 95% CIs to predict the presence of kataegis results in a higher AUC (0.94). Therefore we used the latter model to predict the presence of kataegis in the 998 TCGA samples.

Fig. S3

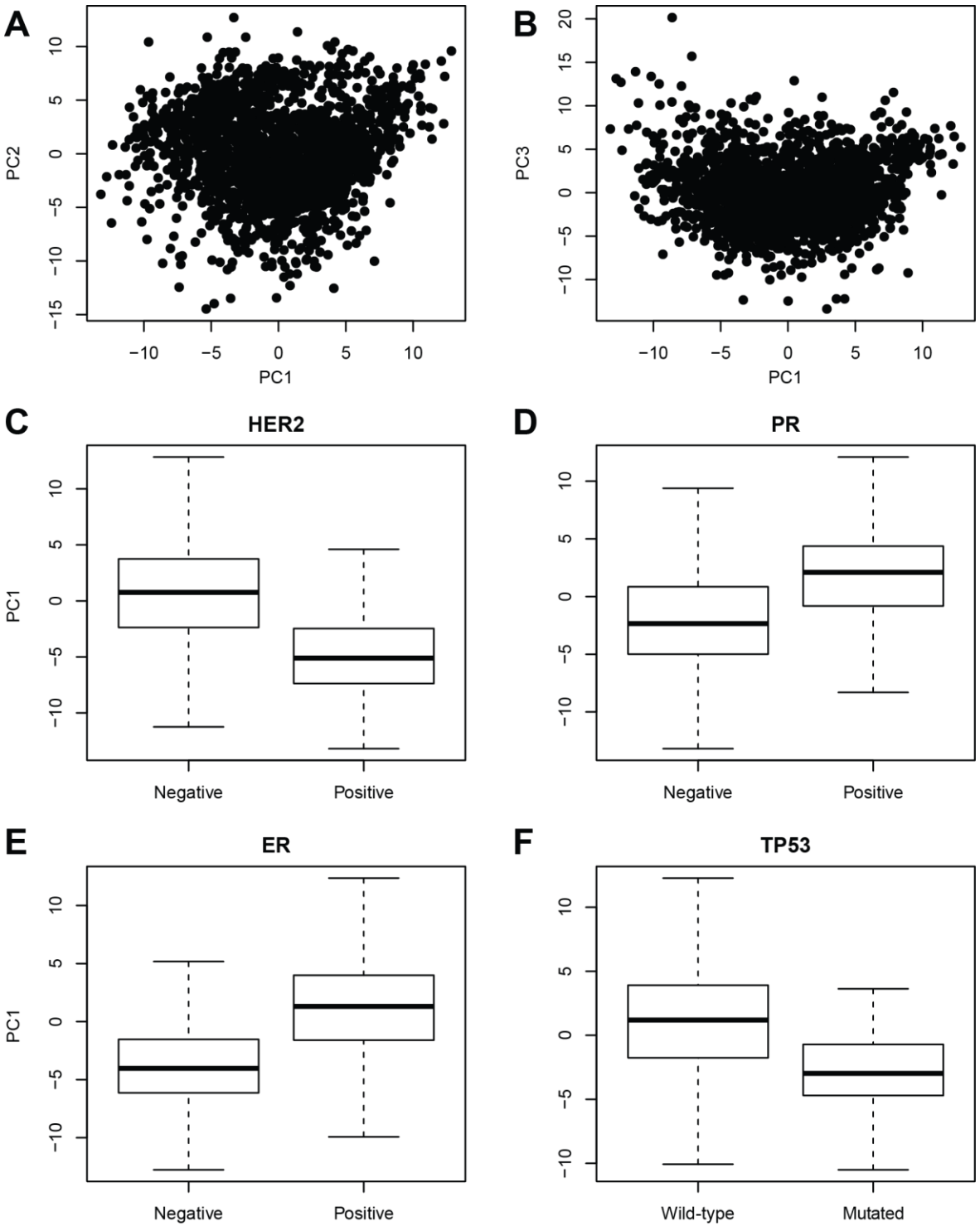


Fig. S3 | Associations between kataegis, hormone receptor status and *TP53* mutations in the METABRIC dataset, related to Fig. 5. To further validate the associations between the kataegis expression signature and clinical features (ER, PR, HER2 status and presence of mutations in *TP53*), we analyzed the METABRIC dataset (Curtis et al., 2012). We

downloaded the expression levels for the 1,992 breast cancer samples in this dataset, as well as the clinical information associated with the corresponding patients. Although METABRIC includes only microarray data, and therefore it is not possible to apply the GLM that we trained in the 97 TCGA discovery tumors to predict the kataegis status, we were still able to examine if the kataegis expression signature is associated with clinical properties.

We performed PCA on the 628 differentially expressed genes in the 97 TCGA discovery tumors. Here, only 555 of these genes have microarray information and were included in the analysis. (**A**) Scatterplot of PC1 and PC2; (**B**) Scatterplot of PC1 and PC3. We investigated the associations between the value of PC1 (which separates presence and absence of kataegis in the original TCGA discovery analysis, Fig. 4A) and (**C**) HER2, (**D**) PR, (**E**) ER status and (**F**) *TP53* mutations. We found that low values of PC1 (associated with presence of kataegis in the 97 TCGA discovery tumors) correspond to ER-negative ($p = 8.3 \times 10^{-28}$, Wilcoxon test), PR-negative ($p = 2.5 \times 10^{-43}$), HER2-positive ($p = 4.4 \times 10^{-37}$), *TP53*-mutated tumors ($p = 2.3 \times 10^{-11}$), similarly to what we observe for the TCGA samples with predicted kataegis status.

Our analysis suggests that the gene expression signature that we observe in the TCGA tumors with kataegis is present in an independent dataset of breast cancer samples.

Fig. S4

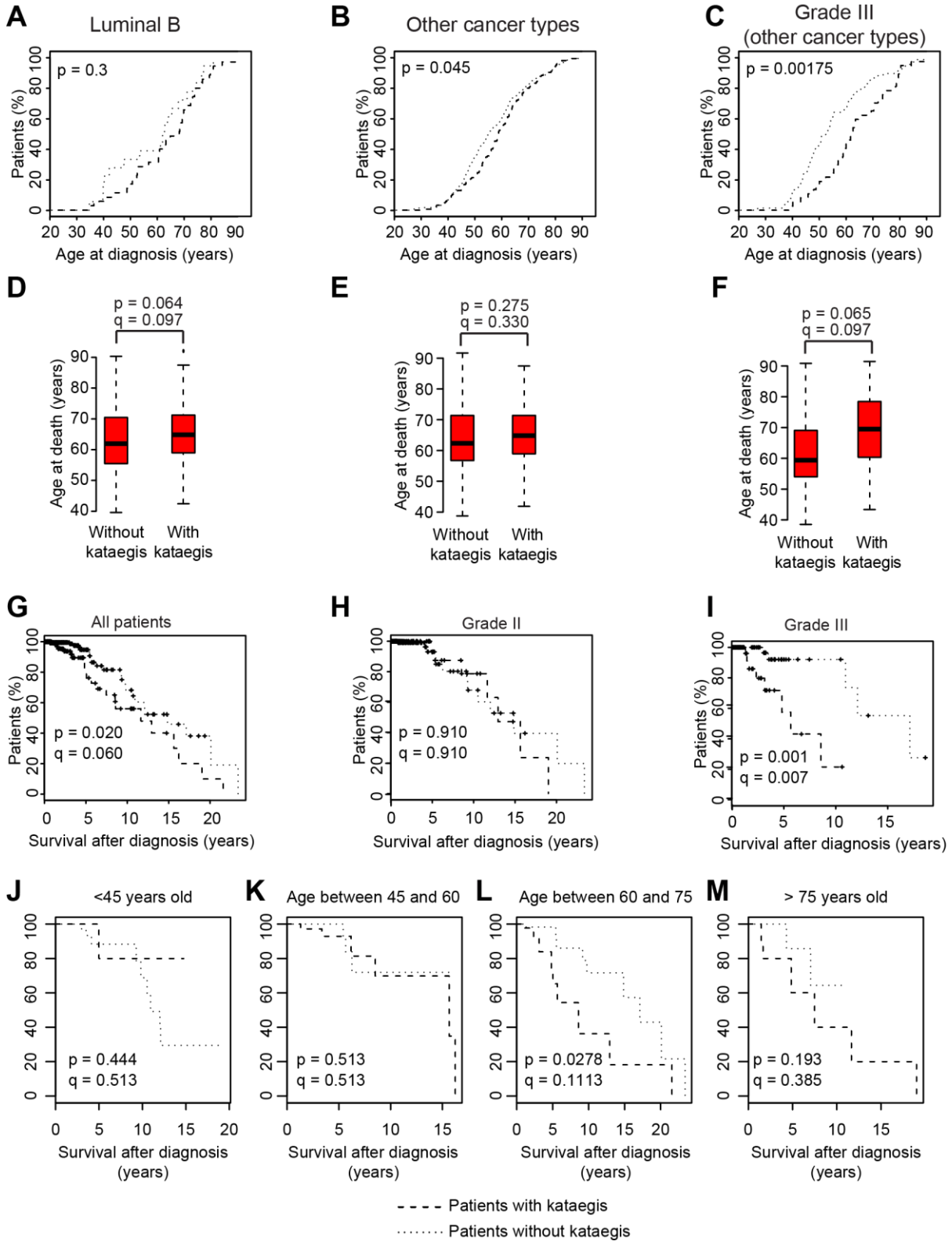


Fig. S4 | Association between predicted kataegis status in 412 TCGA breast cancer patients, diagnosis and survival, related to Fig. 5. (**A-C**) Due to the fact that kataegis loci are enriched in Luminal B tumors we determined if patients with

other tumor types (non-Luminal B) harboring kataegis loci tend to present at a later age. Empirical distributions showing age at diagnosis for **(A)** all patients with Luminal B tumors (n=53), **(B)** all patients with other tumor subtypes (n=359), and **(C)** grade III patients with other tumor subtypes (n=104) with and without kataegis. P-values were calculated using Wald test from Cox proportional hazard model. (**D-F**) Boxplots showing **(D)** the distributions of the age at death (n=45) for all patients, **(E)** for grade II patients (n=20) and **(F)** for grade III patients (n=13) with and without kataegis. (**G-I**) Empirical distributions showing survival after diagnosis for **(G)** all breast cancer patients (n=412), **(H)** patients with grade II tumors (n=277), and (**I**) patients with grade III tumors (n=117). (**J-M**) Survival after diagnosis is shown for patient (**J**) younger than 45 years old (n=80); (**K**) between 45 and 60 (n=194); (**L**) between 60 and 75 (n=155); and (**M**) older than 75 (n=63). P-values in (**A-C, G-M**) were calculated using Wald test from Cox proportional hazard model, while p-values in (**D-F**) were calculated with Wilcoxon test. Benjamini-Hochberg method was used to adjust p-values for multiple testing hypotheses (q-values).

Fig. S5

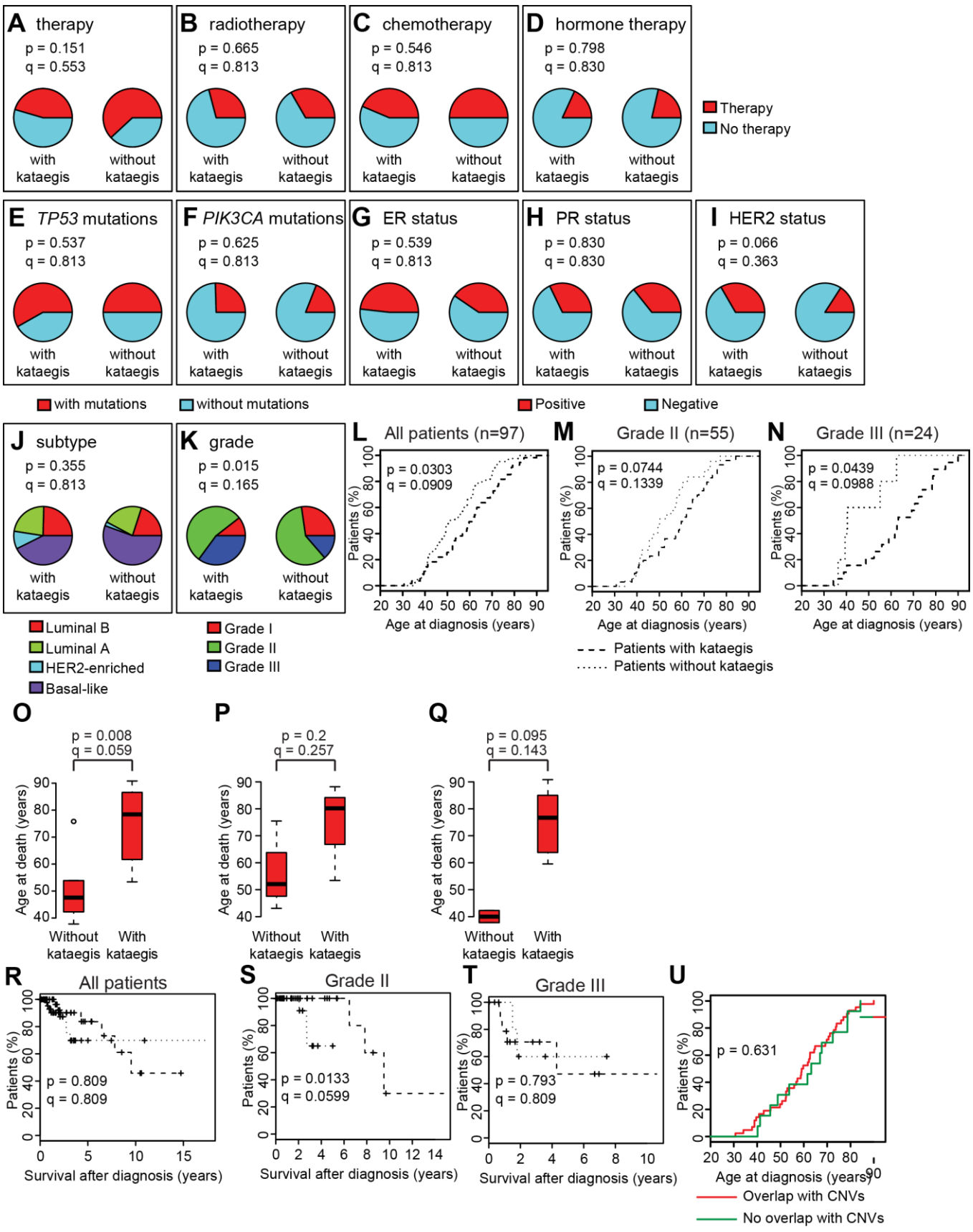

A therapy
p = 0.151
q = 0.553
B radiotherapy
p = 0.665
q = 0.813
C chemotherapy
p = 0.546
q = 0.813
D hormone therapy
p = 0.798
q = 0.830
Therapy
No therapy
with kataegis
without kataegis
E TP53 mutations
p = 0.537
q = 0.813
F PIK3CA mutations
p = 0.625
q = 0.813
G ER status
p = 0.539
q = 0.813
H PR status
p = 0.830
q = 0.830
I HER2 status
p = 0.066
q = 0.363
with mutations
without mutations
Positive
Negative
J subtype
p = 0.355
q = 0.813
K grade
p = 0.015
q = 0.165
Luminal B
Luminal A
HER2-enriched
Basal-like
Grade I
Grade II
Grade III
L All patients (n=97)
p = 0.0303
q = 0.0909
M Grade II (n=55)
p = 0.0744
q = 0.1339
N Grade III (n=24)
p = 0.0439
q = 0.0988
Patients (%)
Age at diagnosis (years)
Patients with kataegis
Patients without kataegis
O
p = 0.008
q = 0.059
P
p = 0.2
q = 0.257
Q
p = 0.095
q = 0.143
Age at death (years)
Without kataegis
With kataegis
R All patients
p = 0.809
q = 0.809
S Grade II
p = 0.0133
q = 0.0599
T Grade III
p = 0.793
q = 0.809
Survival after diagnosis (years)
U
p = 0.631
Overlap with CNVs
No overlap with CNVs

Fig. S5 | Association between kataegis in the 97 TCGA discovery samples and clinical features, related to Fig. 5. The distribution of samples with and without kataegis were examined for differences in: **(A)** Any type of therapy, **(B)** radiotherapy, **(C)** chemotherapy, **(D)** hormone therapy, **(E)** mutations in *TP53*, **(F)** mutations in *PIK3CA*, **(G)** ER status, **(H)** PR status, **(I)** HER2 status, **(J)** tumor subtype and **(K)** grade. Data associated with this figure is shown in Table S1. All data were retrieved from TCGA. The "NA" values in the Table S1 are excluded from these analyses. (**L-N**) Empirical distributions showing **(L)** age at diagnosis (n=97) for all patients, **(M)** for grade II patients (n=55) and **(N)** for grade III patients (n=24) with and without kataegis. P-values were calculated using Wald test from Cox proportional hazard model. P-values were adjusted with Benjamini-Hochberg method (q-values). (**O-Q**) Boxplots showing **(O)** the distributions of the age at death (n=14) for all patients, **(P)** for grade II patients (n=6) and **(Q)** for grade III patients (n=7) with and without predicted kataegis. (**R-T**) Empirical distributions showing survival after diagnosis for **(R)** all breast cancer patients (n=97), **(S)** patients with grade II tumors (n=55), and (**T**) patients with grade III tumors (n=24) for patients with predicted kataegis status. (**U**) Empirical distribution functions showing age at diagnosis for all patients with kataegis loci that overlap (n=42) or do not overlap (n=13) with CNVs. The absence of significant differences in age at diagnosis between samples where kataegis loci overlap and do not overlap CNVs lead us to conclude that CNVs do not influence the association between kataegis loci and age at diagnosis. P-values in (**L-N, R-U**) were calculated using Wald test from Cox proportional hazard model, while p-values in (**O-Q**) were calculated with Wilcoxon test. Benjamini-Hochberg method was used to adjust p-values for multiple testing hypotheses (q-values).

Fig. S6

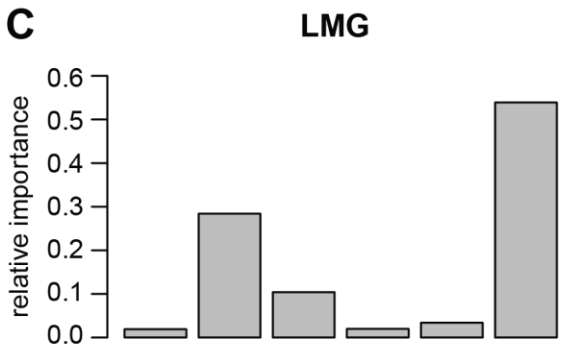
A

Variable	LM coefficients			
	Estimate	Standard Error	t value	p-value (t-test)
(Intercept)	0.280	0.133	2.108	0.038
her2	0.225	0.131	1.718	0.089
pr	-0.230	0.156	-1.474	0.144
er	0.153	0.165	0.929	0.355
tp53	0.063	0.108	0.588	0.558
grade	0.257	0.128	2.002	0.048

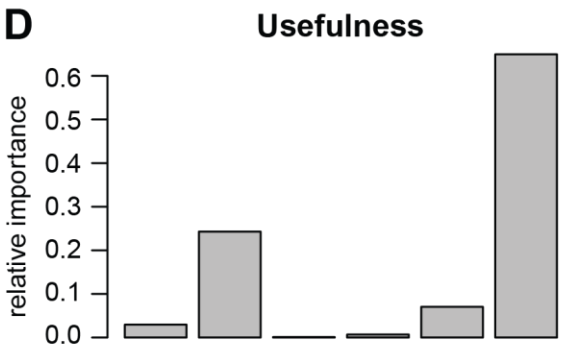
B

Variable	LM coefficients			
	Estimate	Standard Error	t value	p-value (t-test)
(Intercept)	0.110	0.067	1.627	0.107
HER2	0.033	0.066	0.506	0.614
PR	-0.114	0.078	-1.451	0.150
ER	0.008	0.082	0.094	0.926
TP53	0.013	0.053	0.243	0.809
Grade	0.051	0.065	0.781	0.437
Kataegis	-0.123	0.052	-2.372	0.020

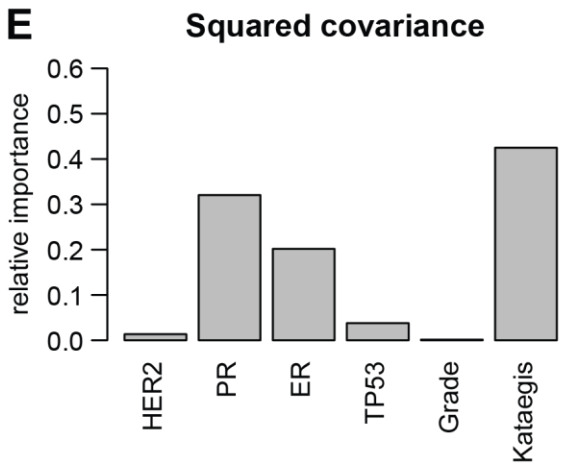
C



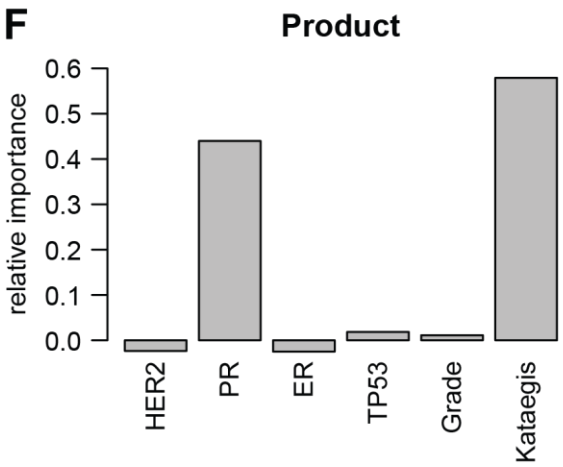
D



E



F



G

Variable	Cox proportional hazard test coefficients				
	Coefficient	Coefficient (exponential)	Standard Error	Z-score	Probability (Z-score)
HER2	-0.110	0.896	0.901	-0.122	0.903
PR	0.091	1.095	1.139	0.080	0.936
ER	-1.264	0.283	1.117	-1.131	0.258
TP53	0.015	1.015	0.697	0.021	0.983
Grade	1.662	5.271	1.103	1.507	0.132
Kataegis	-1.621	0.198	0.761	-2.129	0.033

Fig. S6 | Associations between predicted kataegis status in 412 TCGA tumors and prognosis, related to Fig. 5.

(**A**) To determine the association between kataegis and clinical variables we applied a linear model (LM) using clinical variables as input and predicted kataegis status as output. This analysis shows that only clinical variable that can predict kataegis status is tumor grade. (**B**) To understand the impact of kataegis on prognosis, we applied a linear model using HER2 status, PR status, ER status, presence of mutations in *TP53*, tumor grade and kataegis status as variables to predict prognosis. Bad prognosis is defined as death before age 55. LM coefficients in (**B**) were derived from the linear model output (function lm in R). (**C-F**) Relative importance of each variable in the prediction of prognosis using the LM defined in (**B**) is plotted using four methods included in the *rela.impo* function in R: (**C**) R^2 contribution averaged over orderings among regressors (LMG); (**D**) usefulness; (**E**) squared covariance between prognosis and each variable; and (**F**) product of the standardized coefficient and the correlation (Darlington, 1968; Grömping, 2007). The plots show that kataegis affect the prediction of prognosis more than the other variables, confirming the association between kataegis and death at young age. (**G**) To further confirm the impact of kataegis on prognosis, we developed a Cox proportional hazard model using HER2 status, PR status, ER status, tumor grade, kataegis status and presence of mutations in *TP53* as variables to predict age at death. This table shows that kataegis is the only variable that is significantly associated with age at death, confirming its value as prognostic marker. Cox proportional hazard coefficients were retrieved from the coxph function in the survival package in R.

Fig. S7

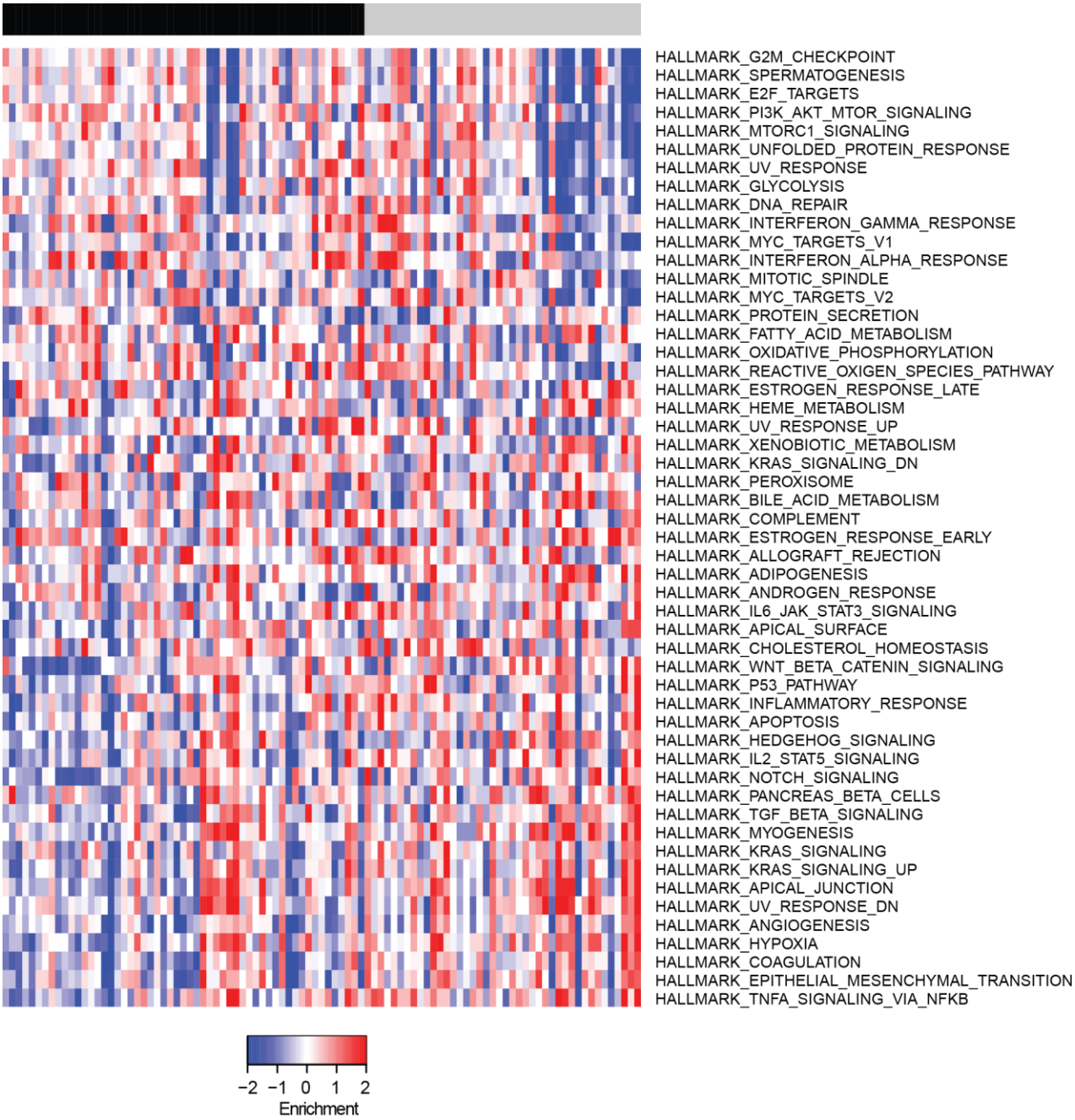


Fig. S7 | Associations between kataegis loci and Hallmark gene sets, related to Figs. 6-7. Heatmap showing normalized pathway enrichment for the Hallmark gene sets derived from MSigDB (Liberzon et al., 2015; Liberzon et al., 2011). Enrichment values were normalized to have mean = 0 and standard deviation = 1 for each pathway. Pathways are sorted from the most upregulated to the most downregulated. Samples with kataegis are indicated by a black bar above the heatmap.

Supplemental Tables

Table S1: Clinical, sequencing information and kataegis loci for all 97 breast tumor samples, related to Fig. 1.

(**Table S1A**) Clinical information was retrieved directly from TCGA. Whole-genome sequencing data was downloaded from CGHub and mutations were detected using MuTect. All mutations in repeats (overlapping RepeatMask track from the UCSC Genome Browser) were not considered for further analysis.

(**Table S1B**) The chromosomal coordinates, length and number of mutations comprising each kataegis loci are given. We also show whether or not the kataegis loci co-localizes with a CNV(s) retrieved from TCGA.

Table S2: Overlap between kataegis loci, chromatin marks and transcription factor binding sites, related to Fig. 2.

Overlap with chromatin marks and transcription factor binding sites was calculated for all kataegis loci. Column C indicates the fraction of kataegis loci overlapping each mark. The position of each kataegis locus was shuffled 10,000 times and for each permutation overlap with all marks and binding sites was calculated. Z-score was calculated as the difference between the observed overlap and the mean over the 10,000 permutations, divided by the standard deviation over the permutations.

Table S3: Expression analysis of samples with and without kataegis, related to Fig. 4.

For each of the 97 breast cancer samples normalized expression data for 20,502 genes was retrieved from TCGA. Mean expression levels were calculated separately between the 55 samples with kataegis and the 42 without kataegis. Tests were performed with edgeR and p-values were adjusted with Benjamini-Hochberg method for multiple testing hypothesis.

Table S4: Prediction of the presence of kataegis in 412 breast cancer samples, related to Fig. 4.

(**Table S4A**) Shown are the predictions of the presence of kataegis using the three methods described in Fig. S4. Column 2 shows the observed value (1 if kataegis is present, 0 if it is absent). GLM includes fit and 95% CI interval. GLM prediction is 0 if the 95% CI is included between 0 and 0.25, 1 if between 0.75 and 1, NA otherwise. The values of

predictions (GLM fit, GLM prediction and random forest prediction) were used to calculate ROC AUC. These values are shown on line 103.

(**Table S4B**) Shown are the 412 TCGA breast cancer samples for which the presence or absence of kataegis loci was predicted. The GLM prediction model was trained to have values = 1 in case of the presence of kataegis loci, and values = 0 in case of absence of kataegis loci. Therefore we defined samples with lower 95% C.I. boundary > 0.75 as predicted to have kataegis, and samples with upper 95% C.I. boundary < 0.25 as predicted not to have kataegis.

Table S5: IC analysis on all tumors with kataegis, related to Fig. 6.

Shown are IC analyses on 10 gene sets. Nominal p-values were adjusted using FDR, FWER and Bonferroni methods. IC analysis was performed on: 1) all tumors with kataegis (columns B-F); 2) tumors with kataegis loci on chromosome 8 (columns G-K); 3) tumors with kataegis loci on chromosome 17 (columns L-P); and 4) tumors with kataegis loci on chromosome 22 (columns Q-U). IC analysis was performed using:

A. Gene expression levels
B. protein levels (RPPA)
C. Chemical and genetic perturbations
D. Biocarta pathways
E. Reactome pathways
F. miRNA targets
G. Transcription factor targets
H. Cancer gene neighborhoods
I. Cancer modules
J. Oncogenic signatures

Table S6: ssGSEA on 50 Hallmark gene sets, related to Fig. 7.

Shown is ssGSEA on the 50 Hallmark gene sets (Figure S7). Enrichment was calculated for each of the 97 breast cancer samples. To determine if any gene set is significantly associated with kataegis, differences in the mean enrichment between samples with and without kataegis were calculated using Information Coefficient analysis (Kim et al., 2016).

References

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., METABRIC Group, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups Nature, 486 (2012), 346–352.

Kim, J.W., Botvinnik, O.B., Abudayyeh, O., Birger, C., Rosenbluh, J., Shrestha, Y., Abazeed, M.E., Hammerman, P.S., DiCara, D., Konieczkowski, D.J., *et al.* (2016). Mapping Genomic Alterations to Functional Profiles of Pathway Activation, Gene Dependency and Drug Sensitivity. Nature biotechnology *in press*.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739-1740.

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1, 417-425.